Linear Scaling 3D Fragment Method for Large-scale Electronic Structure Calculations

Lin-Wang Wang, Zhengji Zhao and Juan Meza

October 16, 2006

Abstract

We present a linear scaling 3 dimensional fragment (LS3DF) method that uses a novel decomposition and patching scheme to do *ab initio* density functional theory (DFT) calculations for large systems. This method cancels out the artificial boundary effects that arise from the spatial decomposition. As a result, the LS3DF results are essentially the same as the original full-system DFT results with errors smaller than the errors introduced by other sources of numerical approximations. In addition, the resulting computational times are thousands of times smaller than conventional DFT methods, making calculations with 100,000 atom systems possible. The LS3DF method is applicable to insulator and semiconductor systems, which covers a current gap in the DOE's materials science code portfolio for large-scale *ab initio* simulations.

1 Introduction

Nanostructures such as quantum dots and wires, composite quantum rods and core/shell structures have been proposed for electronic devices or optical devices like solar cells. Understanding the electronic structures of such systems and the corresponding carrier dynamics is essential to the successful design and deployment of such devices. Yet despite more than a decade of research, some critical issues of the electronic structure of even

moderately complex nanostructures are still poorly understood. One such issue is the internal electric field in a composite colloidal nanostructure and its consequences on the electron wavefunctions. It is well known that there are strong internal electric fields in some bulk semiconductor heterostructures, like the InN/GaN superlattice. These electric fields could be caused by surface and interface dipoles, total dipole of the nanostructure, piezoelectric effects, surface trapped charges or charged dopants. These electric fields can induce strong spatial localizations of the wavefunctions, thus causing different electron-hole recombination rates, charge transports, and nonlinear optical properties, all of which are important to the performance of the nanostructure electro-optical devices. Unfortunately, the continuous model used in conventional device simulations can no longer be used for these nanostructures due to the atomic nature of the charge, dopant and geometry, the high order effects of different phenomenon, and the change of dielectric functions. To address these and other issues, what is needed is a direct atomistic *ab initio* selfconsistent calculation for the charge density and the electric field, and the corresponding atomic relaxation for the nanosystem.

Density functional theory (DFT) is the most widely used *ab initio* method in material science simulations accounting for 75% of the National Energy Research Scientific Computing Center (NERSC) allocation time in the materials science category [1]. DFT codes can be used to calculate the electronic structure, the charge density, the total energy and the atomic forces of a material system, and with the advance of new algorithms and more powerful computers, DFT can now be used to study thousand-atom systems. Unfortunately, conventional DFT algorithms (as implemented in codes like Qbox, Paratec, Petot) scale as $O(N^3)$, where N is the size of the system, putting many problems beyond the reach of even planned petascale computers. While planewave local density approximation (LDA) codes, like Qbox, have demonstrated the capability of using hundreds of thousands of processors on the BlueGene/L computer [2], the $O(N^3)$ scaling for its floating point operation (and $O(N^2)$ communication) may not necessarily be the most efficient way to

solve a given science problem. In addition, as some of the proposed petascale computers envision having millions of processors, new computational paradigms and algorithms are needed to solve these problems efficiently.

To reduce the time to solution, one would ideally like to use a linear scaling *ab initio* method [3]. Most of these methods use localized orbitals, and minimize the total energy as a function of these orbitals. Unfortunately, the use of localized orbitals can introduce extraneous local minima in the total energy functional, which makes the total energy minimization more difficult. On the computational side, it is difficult to effectively use thousands of processors because localized orbitals can have strong overlaps that make parallelization a nontrivial task. As a result of these challenges, the application of these types of linear scaling methods is still quite limited. Another $O(N)$ approach, the LSMS method [4], has also been shown to scale to thousands of processors. However this method can only be applied to metals and has been mainly used to study metallic alloys and magnetic systems. The LSMS method could not be used to study, for example, the class of semiconductor nanostructures discussed above.

For a large materials science problem, a natural approach to achieve linear scaling is by a divide-and-conquer approach: spatially divide the system into many small pieces and solve each piece independently by a small group of processors. The method we present here addresses both the linear scaling and the parallelization issues and could be used to study some of the systems described above. For example, to study the internal electric field of composite nanostructures, we would need to simulate systems with 100,000 atoms. Partly because of the scaling of conventional DFT codes and insufficient computational power, the internal electric field problem has remained as one of the outstanding unsolved problems in colloidal nanoscience. For example, we don't even know whether there is a large internal electric field in a simple quantum dot consisting of polar semiconductors, and if there is one, what is the cause of such an internal electric field. Other examples that could be addressed using our method include: grain boundary, dislocation energies

and atomic structures, impurity transport and clustering in semiconductors, nanostructure growth, and electronic structures of nanostructures.

2 Formalism

Our approach is based on the observation that the total energy of a system can be decomposed into the quantum mechanical part (the wavefunction kinetic energy and the exchange correlation energy), and the classical electrostatic part [5]. While the electrostatic energy (Coulomb energy) is long range, the quantum mechanical energy is local in nature (nearsighted). The long range Coulomb interaction can be solved efficiently using various methods (e.g. Poisson solvers) even for million-atom systems. On the other hand, the quantum mechanical part is more problematic. We will take advantage of the locality of this energy by using a spatial decomposition divide-and-conquer approach. While there are previous methods [6, 7] based on this divide-and-conquer concept, they all rely on positive spatial partition functions to divide and patch the spaces. There are intrinsic difficulties in using this positive partition function technique, in particular for dividing the kinetic energy terms. In contrast, our new division-patching method avoids these problems, resulting in a more accurate algorithm. In fact, we will demonstrate that the results achieved by our method is essentially the same as the original full system LDA method.

Our linear scaling 3D fragment (LS3DF) spatial division-patching technique is inspired by the fragment molecular orbital (FMO) method proposed by Kitaura *et al* [8, 9] and combined with ideas from our own charge patching techniques [10]. FMO is used for organic chain-like molecules, where the long chain molecules are sub-divided into fragment pieces. A full DFT calculation is then done on each piece and pairs of nearby pieces. The total electron charge density is summed over all the pieces and their pairs, with a negative sign for the pieces and a positive sign for the pairs themselves. The use of pairs and negative signs is innovative as this allows the calculation of the energy of the artificial boundaries, which can subsequently be subtracted from the total energy and charge density

summation.

Our LS3DF method also extends the above ideas to 3 dimensional systems and fragments. However, instead of using pairs of pieces, we divide the system using overlapping regions (pieces, fragments). More specifically, our division scheme is illustrated in Fig. 1 for a 2 dimensional system for simplicity. Here, a supercell is divided into $m1 \times m2$ grid points. From each grid point corner (i_1, i_2), we can define four pieces with dimension: $1 \times 1, 1 \times 2, 2 \times 1, 2 \times 2$ respectively. Note that, they are overlapping pieces. Now, after all the pieces at all the (i_1, i_2) corners are calculated, the total charge density of the whole system can be patched together as: $\rho_{tot}(r) = \sum_{(i_1,i_2),D} sign_D \rho_{(i_1,i_2),D}(r)$, where D denotes the dimension $1 \times 1, 1 \times 2, 2 \times 1, 2 \times 2$, and $sign_D$ is $+$ for 1×1 and 2×2, and $-$ for $1 \times 2, 2 \times 1$. The total energy can be expressed in similar fashion using the wavefunctions of each piece, although the electron-electron Coulomb interaction is computed based on the total charge density $\rho_{tot}(r)$.

To illustrate the above formula, we can check each point inside a piece (A point in Fig. 1). Note that each spatial point will be included in 3^2 pieces: four 2×2 pieces, two 2×1 pieces, two 1×2 pieces, and one 1×1 piece. After the above +/- cancellations, it will be covered by only one piece, which is what is needed. We can also check for each boundary point. A boundary can be defined with a direction, that is a boundary from A to B is different than a boundary from B to A; we have used an arrow in Fig. 1 to represent a directional boundary. A given directional boundary is covered by 6 pieces, with equal numbers of positive and negative signs. Since all these pieces have the same (directional) boundary at that point, and given the nearsightedness, their charge density will be the same near that point. As a result, the boundary effects will be canceled out. The same is true for the corner effects. This division scheme can be extended to 3 dimension, where at each corner point (i_1, i_2, i_3), there will be eight pieces: $1 \times 1 \times 1, 1 \times 1 \times 2, 1 \times 2 \times 1, 2 \times 1 \times 1, 1 \times 2 \times 2, 2 \times 1 \times 2, 2 \times 2 \times 1, 2 \times 2 \times 2$. In this case, each spatial point will be covered by 3^3 pieces. The sign in the formula is positive for $2 \times 2 \times 2, 1 \times 1 \times 2, 1 \times 2 \times 1, 2 \times 1 \times 1$,

while negative for $2 \times 2 \times 1, 2 \times 1 \times 2, 1 \times 2 \times 2, 1 \times 1 \times 1$.

The nearsightedness is due to an energy gap between the occupied states and the unoccupied states in a semiconductor system. In order to keep this near-sighted property within each piece, we need to maintain an energy gap within each piece. This can be achieved by a proper surface passivation (usually with H atoms) for the dangling bonds created by the artificial boundaries of the fragments. Each small piece is solved using a conventional planewave code (PEtot [11]). A minor technical detail is that a small area representing a vacuum around the original fragment is added to each fragment prior to using PEtot. When the wavefunctions are solved for each piece, there is no communication needed between the pieces so each piece can be solved independently of all of the other ones. In addition, since the accuracy of this method depends only on the size of the pieces, a large system can attain the same level of accuracy, at the cost of generating more pieces. This makes the total floating point operation scale as $O(N)$ and also makes it easily parallelized to a large number of processors. Thus we believe the LS3DF method is a good candidate for petascale calculations.

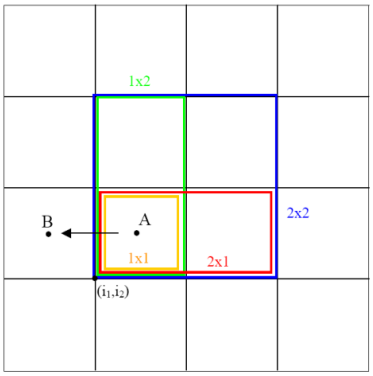

Figure 1: The division of space and fragment pieces from corner (i_1, i_2)

The LDA total energy in the original Kohn-Sham formalism can be expressed as:

$$E_{tot} = \sum_{i=1,M} \int \psi_i^*(r)[-\frac{1}{2}\nabla^2]\psi_i(r)dr + \int V_{ion}(r)\rho_{tot}(r)dr + \frac{1}{2}\int \frac{\rho_{tot}(r)\rho_{tot}(r')}{|r-r'|}drdr' + \int \epsilon_{xc}(\rho_{tot}(r))\rho_{tot}(r)dr, \quad (1)$$

where $\epsilon_{xc}(\rho)$ is the LDA formula for local exchange and correlation energy, ψ_i is the electron wavefunction, and M is the number of occupied states, which is proportional to the number of atoms N in the system. The total charge density $\rho_{tot}(r)$ is defined by

$$\rho_{tot}(r) = \sum_{i=1,M} |\psi_i(r)|^2. \quad (2)$$

For the LS3DF method, the total energy can be expressed as:

$$E_{tot} = \sum_F \alpha_F \sum_i O(\epsilon_{F,i}, E_F) \int \psi_{F,i}^*(r)[-\frac{1}{2}\nabla^2]\psi_{F,i}(r)dr + \int V_{ion}(r)\rho_{tot}(r)dr + \frac{1}{2}\int \frac{\rho_{tot}(r)\rho_{tot}(r')}{|r-r'|}drdr' + \int \epsilon_{xc}(\rho_{tot}(r))\rho_{tot}(r)dr + \sum_F \alpha_F \int \Delta V_F(r)\rho_F(r)dr, \quad (3)$$

where F is the index of the fragment, it is a combined index of (i_1, i_2) and D discussed above (see Fig. 1), α_F is the sign of the fragment, which depends only on $D, O(\epsilon_{F,i}, E_F)$ is an occupation number for the fragment wavefunction $\psi_{F,i}(r)$, depending on its eigenenergy $\epsilon_{F,i}$ and the full space Fermi energy E_F. The total charge density $\rho_{tot}(r)$ is patched from the fragment charge density

$$\rho_{tot}(r) = \sum_F \alpha_F \rho_F(r), \quad (4)$$

while the fragment charge density $\rho_F(r)$ for fragment F is calculated by

$$\rho_F(r) = \sum_i |\psi_{F,i}(r)|^2 O(\epsilon_{F,i}, E_F). \tag{5}$$

Note that both the fragment wavefunction $\psi_{F,i}(r)$ and the fragment charge density $\rho_F(r)$ are only defined within a fragment space (one piece in Fig. 1 plus a small surrounding margin space) Ω_F. The wavefunctions $\psi_{F,i}(r)$ from the same fragment F satisfy the orthonormal condition:

$$\int_{\Omega_F} \psi^*_{F,i}(r)\psi_{F,j}(r)dr = \delta_{i,j}. \tag{6}$$

In Eq. (3) for the LS3DF total energy, the term $\Delta V_F(r)$ represents the surface passivation potential. This additional potential includes the effects of additional H atoms at the artificially created boundaries. This term is only non zero near the artificial surfaces. For fragments that have a common surface, their $\Delta V_F(r)$ will be the same at that surface. As a result, at the boundary points, the last integral in the LS3DF total energy expression at a given surface region B can be re-written as:

$$\int_B \Delta V_F^B(r) \sum_{F'} \alpha_{F'} \rho_{F'}(r) dr,$$

where, F's are the fragments that have the common surface B, and $\Delta V_F^B(r)$ is the common $\Delta V_F(r)$ at the surface B.

Using the arguments we have before for the cancellation of the artificial boundary charge density, $\sum_{F'} \alpha_{F'} \rho_{F'}$ should be zero near the boundary where $\Delta V_F^B(r)$ is non zero. As a result, the total magnitude of this artificially introduced term should be small. Nevertheless, it is important to include this term, because it provides the surface passivation for each fragment and maintains the band gap.

In the LS3DF formalism, the total energy is variational to the fragment wavefunctions; as a result, the fragment wavefunction $\psi_{F,i}(r)$ is the solution of the following fragment

Kohn-Sham equation derived from $\delta E_{tot}/\delta\psi^*_{F,i}(r) = \alpha_F O(\epsilon_{F,i}, E_F)\epsilon_{F,i}\psi_{F,i}(r)$:

$$[-\frac{1}{2}\nabla^2 + V_F(r)]\psi_{F,i}(r) = \epsilon_{F,i}\psi_{F,i}(r), \tag{7}$$

where

$$V_F(r) = V_{tot}(r) + \Delta V_F(r) \quad \text{for} \quad r \in \Omega_F \tag{8}$$

and $V_{tot}(r)$ is the usual LDA total potential calculated from $\rho_{tot}(r)$ by solving the Poisson equation for the electrostatic potential.

3 Numerical Results and Code Scaling

The LS3DF algorithm is extremely accurate when compared to the full-system direct LDA calculation. For example, using a cubic 8 atom cell in a diamond Si structure as our smallest $1 \times 1 \times 1$ piece to calculate a Si quantum dot passivated with H atoms, we found that the relative energy difference between the current method and the direct LDA is 8.E-6, which is smaller than the error introduced by other sources of numerical approximations. The absolute energy difference is less than 8 meV/atom (0.2Kcal/mol). In addition, the total electron charge density has a relative error of 0.02% as shown in Fig. 3, which is essentially the same as the directly calculated results.

Finally, the atomic force error is 6.4E-5 a.u, which is an order of magnitude smaller than the typical stopping criterion used in *ab initio* atomic relaxation. Thus, for all practical purposes, the result of the LS3DF method can be considered the same as the LDA calculation. Also, note that the final accuracy depends on the size of the pieces, so that for the same accuracy, a large system can be divided into similar sized fragments, at the cost of a larger number of fragments. This ensures that the new algorithm scales linearly, has good parallelization properties, and sufficient accuracy.

The LS3DF code is based on the planewave DFT PEtot code [11]. The flow chart of

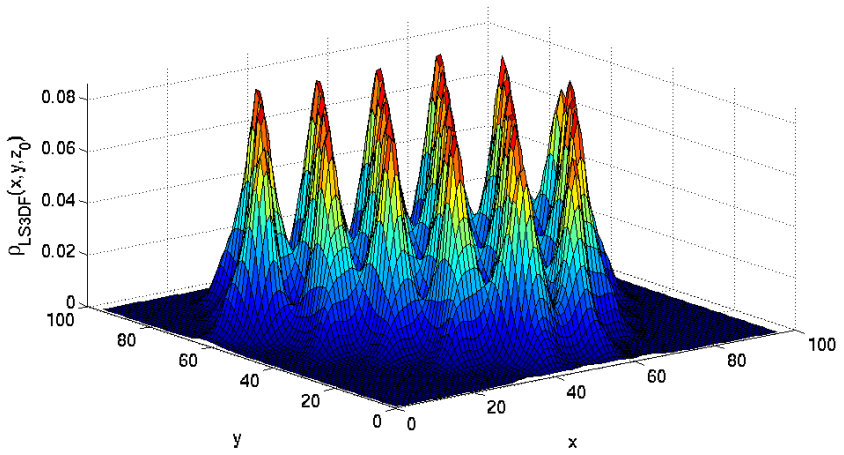

Figure 2: The charge density calculated by the LS3DF method for a Silicon quantum dot, $Si_{235}H_{104}$.

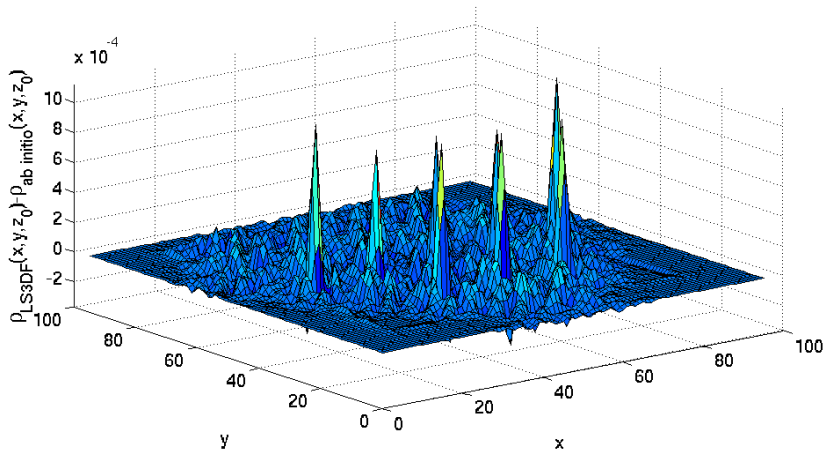

Figure 3: The error of the charge density calculated by the LS3DF method as compared to the direct *ab initio* calculation.

the LS3DF code is shown in Fig. 4 (right hand side), in comparison to the original LDA code (left hand side). The LS3DF code consists of several components: PEtot_F, which divides the number of processors into processor groups, and the calculation of the fragment wavefunctions $\psi_{F,i}(r)$ by each group for a given fragment potential $V_F(r)$ according to Eq. (7). It also calculates the fragment charge density $\rho_F(r)$ from the wavefunction $\psi_{F,i}(r)$ using Eq. (5); Gen_dens patches together the fragment charge densities $\rho_F(r)$ to generate the total charge density $\rho_{tot}(r)$ of the whole system according to Eq. (4). The Poisson step generates the LDA total potential $V_{tot}(r)$ from the total charge density $\rho_{tot}(r)$. This step solves the Poisson equation for the whole system using a global FFT. It also uses the Pulay scheme to mix the resulting LDA potential that is used in the next iteration. Finally, Gen_V_F generates the fragment potential $V_F(r)$ from the input total potential $V_{tot}(r)$ according to Eq. (8).

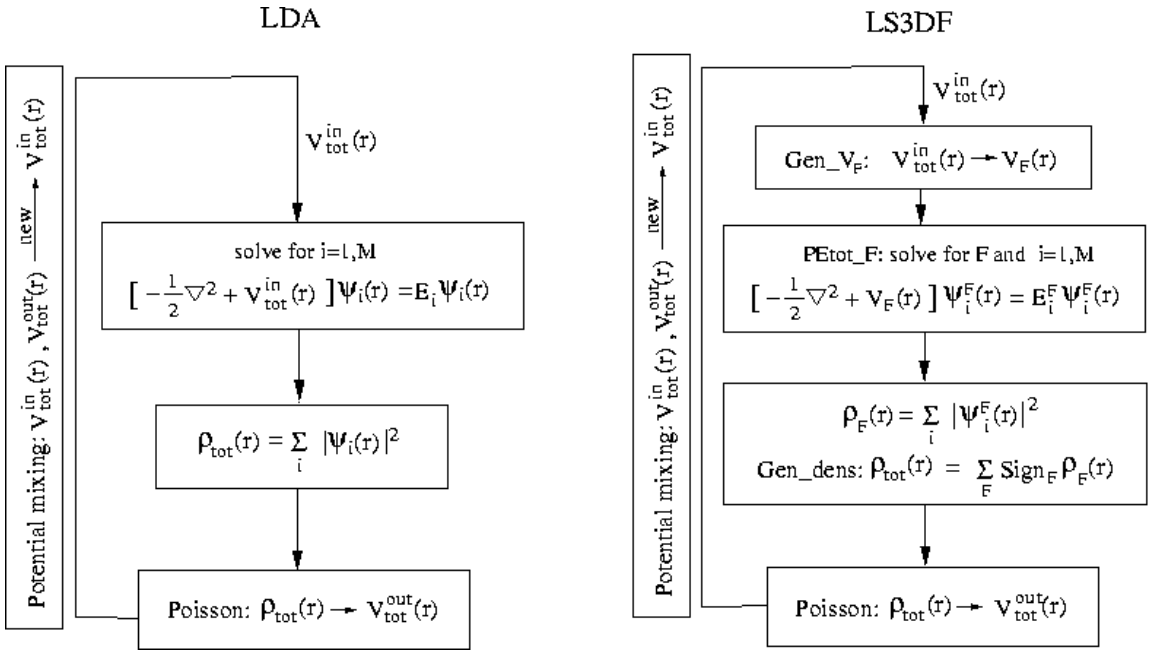

Figure 4: Program flow chart for a conventional LDA method and the LS3DF method

All the codes in Fig. 4 are parallelized. As a demonstration, we have used this approach

to calculate a 3000 atom $Si_{2253}H_{652}$ quantum dot. The calculated total charge density of this quantum dot is shown in Fig. 5. The calculation was done on 1024 processors of the NERSC *Seaborg* computer (IBM RS/6000 SP has 416 16-CPU POWER3+ SMP nodes with a peak performance of 10 teraflop/s).

In Fig. 4, the PEtot_F step is the most time consuming part. The scaling of this part for the 3000 quantum dot (QD) test case on *Seaborg* is shown in Fig. 6. As can be seen, this step scales well up to 1024 processors. We believe it should scale effectively to tens of thousands of processors since each processor group solves the fragments independently.

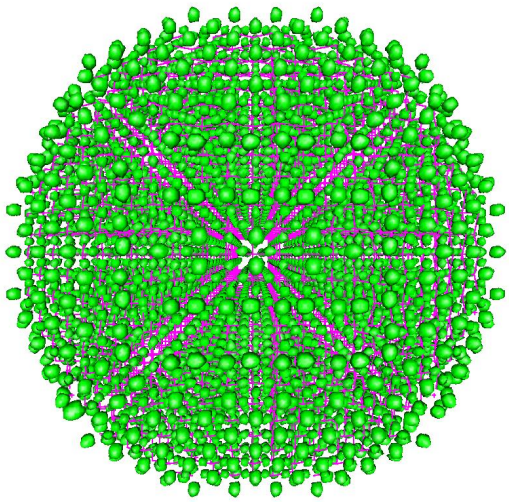

Figure 5: The charge density isosurface (green) plot of a 3000 atom Si quantum dot ($Si_{2253}H_{652}$) passivated by H atoms. The pink color indicates the bonds.

For our 3,000 Si atom QD, when we used 1024 processors, the PEtot_F part of each self-consistent iteration took about 10 minutes. The Poisson solver we used in this calculation was based on FFTs. The real space numerical grid employed was $240 \times 240 \times 240$ and it took about 1 minute to solve the Poisson equation using 128 processors. It took about half a minute each to finish the Gen_dens and Gen_V_F program using 128 processors. Thus,

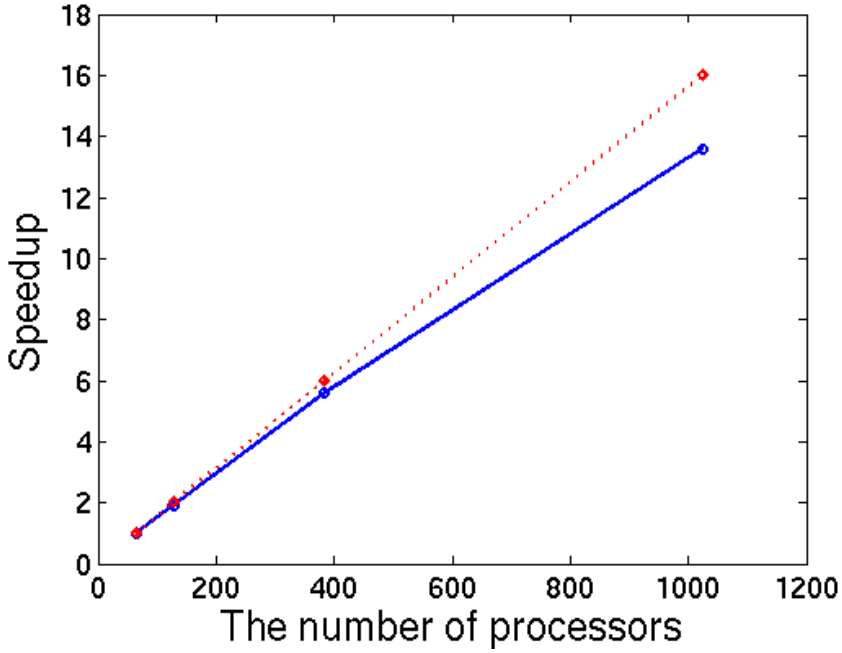

Figure 6: The speedup (blue line) as a function of the number of processors used in PEtot_F.

in total, for the 1024 processors calculation above, it took approximately 12 minutes to finish one self-consistent field (SCF) step. For such a large QD, it typically takes 10 to 20 SCF steps (the outer loop in Fig. 4) to converge the self-consistent iteration. As a result, it will take about 2 to 4 hours to finish one self-consistent calculation (for a fixed atomic position) for this type of QD system using 1024 processors.

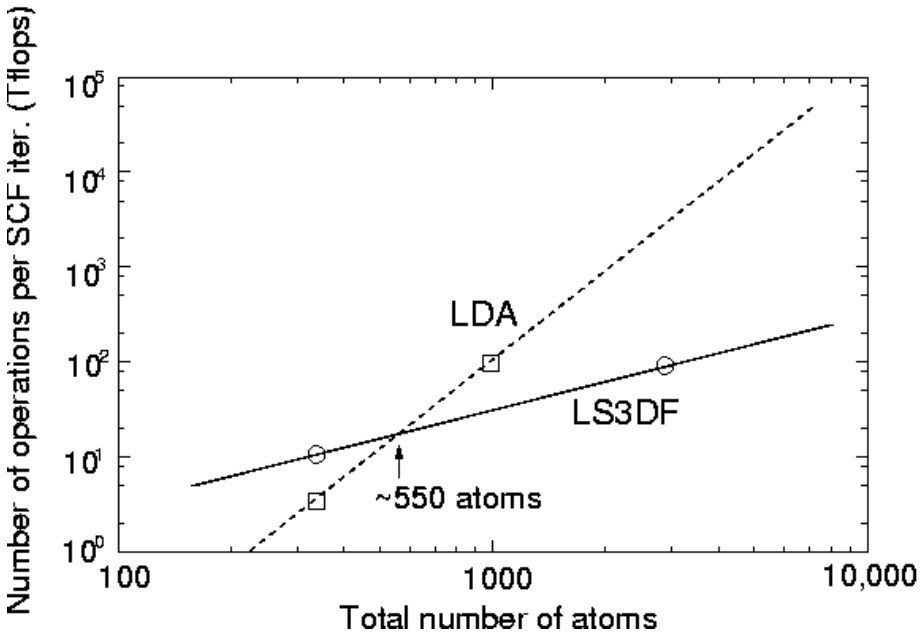

Figure 7: Total floating point operation per SCF iteration for LS3DF method and LDA, which was obtained by using NERSC's profiling tool *ipm*.

The comparison between the $O(N)$ LS3DF with a conventional $O(N^3)$ LDA method is shown in Fig. 7 for the total number of floating point operations for one SCF step. As we can see, the cross over based on the floating point operation counts is at about 500 atoms, which is similar to the reported cross over for the localized orbital method [12].

Acknowledgments This wok was supported by the Director, Office of Science, Basic Energy Sciences, and Division of Material Science, and the Advanced Scientific Computing Research office, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. It used the resources at the National Energy Research Scientific Computing Center (NERSC).

References

[1] L.-W. Wang, A survey of codes and algorithms used in NERSC material science allocations, *LBNL report 61051*, Lawrence Berkeley National Laboratory.

[2] ACM, *Large-Scale First-Principles Molecular Dynamics Simulations on the BlueGene/L Platform using the Qbox Code*, 2005.

[3] G. Goedecker, Linear scaling electronic structure methods. *Rev. Mod. Phys*, 71:1085, 1999.

[4] PSC, *Locally Self-Consistent Multiple Scattering (LSMS) Method*, http://www.psc.edu/general/software/packages/lsms/.

[5] W. Kohn, Density functional and density matrix method scaling linearly with the number of atoms, *Phys. Rev. Lett.*, 76(17):3168–3171, 1996.

[6] W. Yang, Direct calculation of electron density in density-functional theory, *Phys. Rev. Lett.*, 66:1438, 1991.

[7] F. Shimojo, R.K. Kalia, A. Nakano and P. Vashishta, *Comp. Phys. Commun.*, 167:151, 2005.

[8] K. Kitaura, E. Ikeo, T. Asada, T. Nakano, and M. Uebayasi, Fragment molecular orbital method: an approximate computational method for large molecules, *Chem. Phys. Lett.*, 313:701, 1999.

[9] K. Kitaura, S.-I. Sugiki, T. Nakano, Y. Komeiji, and M. Uebayasi, Fragment molecular orbital method: analytical energy gradients, *Chem. Phy. Lett*, 336:163, 2001.

[10] L.-W. Wang and J. Li, First-principles thousand-atoms quantum dot calculations, *Phys. Rev. B*, 69:153302, 2004.

[11] L.-W. Wang, Parallel planewave pseudopotential ab initio package, 2004, http://hpcrd.lbl.gov/ linwang/PEtot/PEtot.html.

[12] J.-L. Fattebert and F. Gygi, Linear-scaling first-principles molecular dynamics with plane-waves accuracy, *Phys. Rev. B*, 73:115124, 2006.